

Optimizing Fronthaul Quantization for Flexible User Load in Cell-Free Massive MIMO

Fabian Götsch^{*†}, Max Franke^{*}, Arash Pourdamghani^{*}, Giuseppe Caire^{*†} and Stefan Schmid^{*}

^{*}Technical University Berlin, Berlin, Germany

[†]Massive Beams, Berlin, Germany

E-mail: {fabian.goetsch, mfranke, pourdamghani, caire, stefan.schmid}@tu-berlin.de

Abstract—We investigate the physical layer (PHY) spectral efficiency and fronthaul network load of a scalable user-centric cell-free massive MIMO system. Each user-centric cluster processor responsible for cluster-level signal processing is located at one of multiple decentralized units (DUs). Thus, the radio units in the cluster must exchange data with the corresponding DU over the fronthaul. Because the fronthaul links have limited capacity, this data must be quantized before it is sent over the fronthaul. We consider a routed fronthaul network, where the cluster processor placement and fronthaul traffic routing are jointly optimized with a mixed-integer linear program. For different numbers of users in the network, we investigate the effect of fronthaul quantization rates, a system parameter computed based on rate-distortion theory. Our results show that with optimized quantization rates, the fronthaul load is quite stable for a wide range of user loads without significant PHY performance loss. This demonstrates that the cell-free massive MIMO PHY and fronthaul network are resilient to varying user densities.

Index Terms—Cell-free massive MIMO, fronthaul, quantization, O-RAN.

I. INTRODUCTION

Cell-free massive MIMO is one of the promising technologies to meet the demands of future 6G wireless communications. As each user equipment (UE) is served by the joint processing of spatially distributed infrastructure antennas, i.e., a user-centric cluster of radio units (RUs), the effects of pathloss and blocking are mitigated, while the macrodiversity is increased and cell boundaries are removed. If the system is well designed and each UE can be served by every RU with a significant channel gain, this design results in a system with highly reduced interference compared to cellular networks.

While the physical layer (PHY) has been investigated in countless works, the fronthaul has only recently been considered in few works (see [1] and references therein). A scalable fronthaul network is considered in [2], [3], where each user-centric cluster processor is located at one of N decentralized units (DUs).¹ In [2], [3], the maximum fronthaul link load is minimized by routing the fronthaul traffic and placing the cluster processors at one of the N DUs. The traffic routing is done via dedicated routers placed in the fronthaul network [2] or routers collocated at DUs, utilizing inter-DU links [3]. In these works, a constant user load is considered.

¹We use the scalability definition of [4], where a system is considered scalable if the communication and computation complexity at each network component remains finite if the network area grows infinitely with constant density of UEs, RUs and DUs.

In a practical network however, the number of UEs in the network can vary drastically, e.g., on a university campus or at a sports venue. By scheduling K active UEs out of the total number K_{tot} of UEs in the network on time-frequency resources, the long-term average throughput rate and idle time of each UE can be optimized. Depending on the user demands, it may be more beneficial to serve many users simultaneously at lower data rates than to serve fewer users at high data rates but with longer idle times. In this case, even with optimized fronthaul routing and cluster placement, increased fronthaul link load is expected due to the larger number of simultaneously active UEs. However, another possible optimization variable to optimize the fronthaul load is the distortion level D with regard to the quantization accuracy.

Fronthaul network planners are generally interested in two metrics: the expected average fronthaul load and the worst-case fronthaul load at peak data traffic events (i.e., very high user load). The average load is important because designing and deploying networks only around peak traffic is costly and will leave vast amounts of unused capacity most of the time [5]. However, it is also crucial to understand what peak traffic events look like due to an increase of demanded user data rates or simultaneously active users in the network. Then, even PHY data rates for the most basic functions such as text messages or calls are hard to achieve, e.g., during sports games or emergencies [6]. Understanding peak traffic events allows operators to design the fronthaul and RAN to be resilient so that they do not collapse under such conditions.

For these reasons, it is important to understand how a relatively small number of users with high data rates will affect the fronthaul network compared to many users with lower data rates. Especially for user-centric cell-free massive MIMO with a routed fronthaul and distributed DUs and RUs, this is a complex and not well-studied problem.

A. Contributions

This paper contributes a first study of how the number of simultaneously active UEs K affects the fronthaul load. This is relevant as it will enable network planners to make RANs more resilient during peak data traffic events. We identify the user load, for which the RAN operates at full PHY performance, and optimize the fronthaul quantization distortion level D . Our results show that the fronthaul load for high user loads is virtually the same as in a lightly loaded system. Further, the

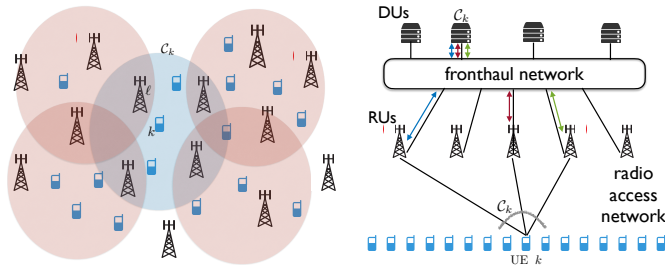


Fig. 1: An example of user-centric clusters, UE-RU association and the data exchange over the fronthaul for user k and its cluster C_k .

PHY spectral efficiency (SE) is degraded only very slightly for optimized D with regard to the fronthaul compared to smaller distortion levels.

II. SYSTEM MODEL

We consider a cell-free massive MIMO RAN and a fronthaul network with a set of Q routers $\mathcal{Q} = \{1, \dots, Q\}$ and N DUs $\mathcal{N} = \{1, \dots, N\}$, respectively. The RAN consists of K single-antenna UEs $\mathcal{K} = \{1, \dots, K\}$ and L RUs $\mathcal{L} = \{1, \dots, L\}$, where each RU is equipped with M antennas. The UE-RU associations in the RAN are described by a bipartite graph $\mathcal{G}_{\text{ran}}(\mathcal{K}, \mathcal{L}, \mathcal{E}_{\text{ran}})$ with two classes of nodes (UEs and RUs). The user-centric RU cluster serving user $k \in \mathcal{K}$ is denoted by $C_k \subseteq \mathcal{L}$. The set of UEs served by RU $\ell \in \mathcal{L}$ is denoted by $\mathcal{U}_\ell \subseteq \mathcal{K}$. The set of edges of \mathcal{E}_{ran} is such that $(\ell, k) \in \mathcal{E}_{\text{ran}}$ iff $\ell \in C_k$ (or, equivalently, iff $k \in \mathcal{U}_\ell$). The fronthaul network connects each RU with a subset of the routers, which are partially connected to the other routers and DUs. It is described as a graph $\mathcal{G}_{\text{front}}(\mathcal{L}, \mathcal{Q}, \mathcal{N}, \mathcal{E}_{\text{front}})$, where the edges $\mathcal{E}_{\text{front}}$ represent the connections between RUs, routers and DUs. The overall network topology is described by the union of $\mathcal{G}_{\text{ran}}(\mathcal{K}, \mathcal{L}, \mathcal{E}_{\text{ran}})$ and $\mathcal{G}_{\text{front}}(\mathcal{L}, \mathcal{Q}, \mathcal{N}, \mathcal{E}_{\text{front}})$ obtained by merging the common nodes \mathcal{L} (see Fig. 1).

The matrix $\mathbb{H} \in \mathbb{C}^{LM \times K}$ describes the channels between the K UEs and LM RU antennas. It consists of the vectors $\mathbf{h}_{\ell,k} \in \mathbb{C}^{M \times 1}$ representing the channel between UE k and RU ℓ . Each channel is a correlated complex circularly symmetric Gaussian vector with mean zero and covariance matrix $\Sigma_{\ell,k} = \mathbb{E}[\mathbf{h}_{\ell,k} \mathbf{h}_{\ell,k}^H]$, where we use the well-known notation $\mathbf{h}_{\ell,k} \sim \mathcal{CN}(\mathbf{0}, \Sigma_{\ell,k})$. The large-scale fading coefficient (LSFC) corresponding to $\mathbf{h}_{\ell,k}$ is defined as $\beta_{\ell,k} \triangleq \frac{1}{M} \text{tr}(\Sigma_{\ell,k})$. It describes the average signal attenuation between RU ℓ and user k due to distance and other macroscopic effects. The channel coefficients remain constant over blocks of T signal dimensions (time-frequency channel uses), referred to as “resource blocks” (RBs). This is a standard model in the (cell-free) massive MIMO literature (see [4], [7]). For simplicity, we focus on a generic resource block and we omit the resource block index in this paper.

A. Cluster and pilot allocation

The user-centric clusters and uplink (UL) pilots for channel estimation are assigned following the subspace information aided overloaded pilot assignment scheme from [8]. All UEs transmit with the same average energy per symbol P^{ue} and we

define the system parameter $\text{SNR} \triangleq \frac{P^{\text{ue}}}{N_0}$, where N_0 denotes the complex baseband noise power spectral density. Each UE is associated to a cluster C_k of RUs of cardinality at most C_{max} (a system parameter) with the largest LSFC, provided that

$$\beta_{\ell,k} \geq \frac{\eta}{\text{MSNR}}, \quad (1)$$

where $\eta > 0$ is a suitable threshold that defines how much larger the useful signal in the presence of maximum possible beamforming gain (equal to M) should be compared to the noise floor. We assume that τ_p out of T signal dimensions per RB are dedicated to pilots and use a codebook of τ_p orthogonal pilot sequences $\{\phi_t : t \in [\tau_p]\}$. Let $\mathbf{F}_{\ell,k}$ denote the matrix of the orthonormal eigenvectors of $\Sigma_{\ell,k}$ spanning the *dominant channel subspace*, i.e., the subspace containing a sufficiently large fraction of the total channel energy $\text{tr}(\Sigma_{\ell,k}) = M\beta_{\ell,k}$ (e.g., see [9], [10]). A RU can assign a pilot to multiple UEs under the condition that the user channel subspaces are approximately mutually orthogonal [2].

We let t_i denote the pilot index of UE i . Then, each RU ℓ obtains an estimate of the channel vectors $\{\mathbf{h}_{\ell,k} : k \in \mathcal{U}_\ell\}$ using the subspace projection method of [9], given by

$$\hat{\mathbf{h}}_{\ell,k} = \mathbf{h}_{\ell,k} + \mathbf{F}_{\ell,k} \mathbf{F}_{\ell,k}^H \left(\sum_{i \neq k: t_i = t_k} \mathbf{h}_{\ell,i} \right) + \mathbf{F}_{\ell,k} \mathbf{F}_{\ell,k}^H \tilde{\mathbf{z}}_{\ell}^{(t)},$$

where $\tilde{\mathbf{z}}_{\ell}^{(t)}$ is projected additive white Gaussian noise (AWGN) with i.i.d. components $\sim \mathcal{CN}(0, \frac{1}{\tau_p \text{SNR}})$ and the sum in the second term contains the channels of co-pilot users with respect to UE k . Since orthogonal pilots are used, the contributions of non-co-pilot users are removed from the received pilot signal by multiplying with UE k 's pilot sequence. When $\mathbf{F}_{\ell,k}$ and $\mathbf{F}_{\ell,i}$ are nearly mutually orthogonal, i.e. $\mathbf{F}_{\ell,k}^H \mathbf{F}_{\ell,i} \approx \mathbf{0}$, the subspace projection is able to significantly reduce the pilot contamination effect. Since the subspace estimation scheme in [9] achieves virtually the same performance as perfect channel subspace knowledge, we assume channel subspaces to be known in this paper.

B. Uplink data rates with fronthaul quantization

Let s_k^{ul} be the UL data symbol of UE k (i.i.d. with mean zero and unit variance) at a given channel use of a generic RB. In this work, we consider “smart RUs” that operate according to O-RAN split option 7.2x in the upstream direction. More specifically, our chosen split option is most similar to 7.2x Cat-B ULPI-A, where RUs do channel estimation and equalization locally [11]. As in [2], [9], we consider “local detection with cluster-level combining”, so that the equalized symbol (i.e., a local observation $r_{\ell,k}^{\text{ul}}$ of s_k^{ul}) at RU ℓ for each UE $k \in \mathcal{U}_\ell$ is sent to the cluster processor of C_k located at some DU over the fronthaul. The cluster-level symbol estimate is then computed at the DU.

In particular, the received UL signal at RU ℓ is given by

$$\mathbf{y}_{\ell}^{\text{ul}} = \sqrt{\text{SNR}} \sum_{i \in \mathcal{K}} \mathbf{h}_{\ell,i} s_i^{\text{ul}} + \mathbf{z}_{\ell}^{\text{ul}}, \quad (2)$$

where $\mathbf{z}_\ell^{\text{ul}}$ is AWGN with components $\sim \mathcal{CN}(0, 1)$. The local observation is obtained as a linear projection of $\mathbf{y}_\ell^{\text{ul}}$ onto a suitably defined receiver vector $\mathbf{v}_{\ell,k}$, such that

$$r_{\ell,k}^{\text{ul}} = \mathbf{v}_{\ell,k}^H \mathbf{y}_\ell^{\text{ul}}. \quad (3)$$

The receiver vectors $\{\mathbf{v}_{\ell,k} : k \in \mathcal{U}_\ell\}$ are computed at RU ℓ using the local channel estimates $\{\hat{\mathbf{h}}_{\ell,k} : k \in \mathcal{U}_\ell\}$. We consider linear MMSE receivers based on the partial local CSI at RU ℓ , given by

$$\mathbf{v}_{\ell,k} = \left(\nu_\ell \mathbf{I} + \text{SNR} \sum_{i \in \mathcal{U}_\ell} \hat{\mathbf{h}}_{\ell,i} \hat{\mathbf{h}}_{\ell,i}^H \right)^{-1} \hat{\mathbf{h}}_{\ell,k}, \quad (4)$$

where $\nu_\ell \triangleq 1 + \text{SNR} \sum_{i \in \mathcal{U}_\ell} \beta_{\ell,i}$ accounts for unknown interference and noise [9]. The cluster processor for user k collects the local observations from all RUs $\ell \in \mathcal{C}_k$ and forms a cluster-level combined observation which is then passed to the channel decoder for user k as the soft-output of a virtual single user additive noise channel.

Since the fronthaul network has finite capacity links, we let each RU quantize its local observation $r_{\ell,k}^{\text{ul}}$ with $B_{\ell,k}$ bits per sample before sending it to the cluster processor. We use the quantization scheme from [2] based on rate-distortion theory, where $B_{\ell,k}$ depends on the signal strength at RU ℓ regarding UE k defined as $\sigma_{\ell,k}^2 \triangleq \mathbb{E}[|r_{\ell,k}^{\text{ul}}|^2]$. For a given desired distortion level D , each RU $\ell \in \mathcal{C}_k$ uses quantization rate

$$B_{\ell,k} = \max \left\{ \log_2 \frac{\sigma_{\ell,k}^2}{D}, 0 \right\} \quad (5)$$

to send the $r_{\ell,k}^{\text{ul}}$ to the corresponding cluster processor. Note that if $\sigma_{\ell,k}^2 < D$ for some $(\ell, k) \in \mathcal{E}_{\text{ran}}$, then $r_{\ell,k}^{\text{ul}}$ is simply not sent from the RU to the cluster processor since the quantization rate is $B_{\ell,k} = 0$. This is equivalent to removing RU ℓ from cluster \mathcal{C}_k . The quantized local observation is given by [2]:

$$\hat{r}_{\ell,k}^{\text{ul}} = \alpha_{\ell,k} r_{\ell,k}^{\text{ul}} + e_{\ell,k}, \quad (6)$$

where

$$\alpha_{\ell,k} = \frac{\sigma_{\ell,k}^2 - D}{\sigma_{\ell,k}^2}, \quad (7)$$

and where $e_{\ell,k}$ is a zero-mean Gaussian random variable uncorrelated with $r_{\ell,k}^{\text{ul}}$ and with variance

$$\hat{\sigma}_{\ell,k}^2 = (1 - D/\sigma_{\ell,k}^2)D. \quad (8)$$

We let $\hat{\mathbf{r}}_k^{\text{ul}} \in \mathbb{C}^{|\mathcal{C}_k| \times 1}$ denote the vector of the quantized local observations $\{\hat{r}_{\ell,k}^{\text{ul}} : \ell \in \mathcal{C}_k\}$. The cluster-level combined observation is then given by

$$\mathbf{r}_k^{\text{ul}} = \mathbf{w}_k^H \hat{\mathbf{r}}_k^{\text{ul}}, \quad (9)$$

where \mathbf{w}_k with elements $\{w_{\ell,k} : \ell \in \mathcal{C}_k\}$ is the cluster-level combining vector that weighs the different local observations. It is computed to maximize the *nominal* Signal to Interference plus Noise Ratio (SINR) of the channel with input s_k^{ul} and output r_k^{ul} , given the local knowledge of the cluster processor \mathcal{C}_k . Due to the limited local knowledge of cluster processor

\mathcal{C}_k , we differentiate between the actual and nominal SINR. The nominal SINR is only used to compute \mathbf{w}_k , while the actual SINR is used to compute the actual achievable physical layer user rate. The maximization of the nominal SINR with respect to the vector of combining coefficients \mathbf{w}_k is a classical generalized Rayleigh quotient maximization problem and is given in [2].

The resulting *actual* SINR conditioned on the realization of all the channel vectors is given by [2]

$$\text{SINR}_k^{\text{ul}} = \frac{\text{SNR} \left| \sum_{\ell \in \mathcal{C}_k} \tilde{g}_{\ell,k,k} \right|^2}{\sum_{\ell \in \mathcal{C}_k} \tilde{d}_{\ell,k} + \text{SNR} \sum_{i \neq k} \left| \sum_{\ell \in \mathcal{C}_k} \tilde{g}_{\ell,k,i} \right|^2},$$

where $\tilde{g}_{\ell,k,i} \triangleq w_{\ell,k}^* \alpha_{\ell,k} \mathbf{v}_{\ell,k}^H \mathbf{h}_{\ell,i}$ and $\tilde{d}_{\ell,k} \triangleq |w_{\ell,k}|^2 \left(\alpha_{\ell,k}^2 \|\mathbf{v}_{\ell,k}\|^2 + \hat{\sigma}_{\ell,k}^2 \right)$. As a performance measure of the physical layer, we consider the UL *Optimistic Ergodic Rate* (OER) [12] given by

$$R_k^{\text{ul}} = \mathbb{E}[\log(1 + \text{SINR}_k^{\text{ul}})], \quad (10)$$

where the expectation is with respect to the small-scale fading, for given values of the LSFCs that depend on the placement of UEs and RUs, the pathloss function and cluster formation.

C. Downlink data rates with fronthaul quantization

In the downlink (DL), the RUs operate according to split option 7.3 of 3GPP (e.g., see [13]), where the cluster processor sends the information bits to the RU, which carries out the modulation and precoding of the data symbols. Since each RU $\ell \in \mathcal{C}_k$ sends the same signal to UE k , the cluster processor sends the same information bits over the fronthaul to the RUs. The DL fronthaul traffic is thus of type multiple-multicast, requiring routers that enable multicast routing. Since the information bits are sent, no quantization is necessary.²

We use the approximate UL-DL reciprocity of [2], [9] for the DL precoding vectors $\mathbf{u}_{\ell,k}$. We let $\mathbf{u}_{\ell,k} \propto w_{\ell,k}^0 \mathbf{v}_{\ell,k}$ where $\mathbf{v}_{\ell,k}$ are the local linear MMSE combiners for the UL defined in (4) and $w_{\ell,k}^0$ are the cluster-level combining coefficients as in (9) for the case of zero quantization distortion. The actual local precoding vectors \mathbf{u}_k are obtained by stacking the blocks $w_{\ell,k}^0 \mathbf{v}_{\ell,k}$ for all $\ell \in \mathcal{C}_k$ and all-zero blocks for $\ell \notin \mathcal{C}_k$, and normalizing the resulting vector of dimension $LM \times 1$ to have unit norm. In this way, except for the common normalization factor, $\mathbf{u}_{\ell,k}$ can be calculated from the local CSI at RU ℓ , with sufficiently high resolution finite arithmetic.

We let \mathbb{h}_k denote the k -th column of the overall channel matrix \mathbb{H} and s_k^{dl} the (coded) information symbol for UE k (independent with mean zero and variance $q_k \geq 0$). The received signal sample at UE k corresponding to one DL channel use is given by

$$y_k^{\text{dl}} = \mathbb{h}_k^H \mathbf{u}_k s_k^{\text{dl}} + \sum_{j \neq k} \mathbb{h}_k^H \mathbf{u}_j s_j^{\text{dl}} + z_k^{\text{dl}}, \quad (11)$$

²However, note that the choice of D still affects the DL physical layer rates. If $\sigma_{\ell,k}^2 < D$ for some $(\ell, k) \in \mathcal{E}_{\text{ran}}$, the corresponding RU is removed from \mathcal{C}_k .

where, without loss of generality, we scale the received signal such that the noise is $z_k^{\text{dl}} \sim \mathcal{CN}(0, \text{SNR}^{-1})$. The corresponding DL SINR is given by [2]

$$\text{SINR}_k^{\text{dl}} = \frac{|\mathbf{h}_k^{\text{H}} \mathbf{u}_k|^2 q_k}{\text{SNR}^{-1} + \sum_{j \neq k} |\mathbf{h}_k^{\text{H}} \mathbf{u}_j|^2 q_j}. \quad (12)$$

We choose uniform power allocation to all data streams, i.e., $q_k = 1$ for all $k \in \mathcal{K}$. The corresponding DL OER is given by

$$R_k^{\text{dl}} = \mathbb{E}[\log(1 + \text{SINR}_k^{\text{dl}})]. \quad (13)$$

Since the information bits are directly sent from the cluster processor located at some DU to the RUs, the fronthaul load corresponding to the DL is equal to the physical layer user rate, i.e., R_k^{dl} .

D. Remarks on the fronthaul load

The choice of the split options aims to reduce the fronthaul load by allocating many PHY functions to the RUs, thereby limiting the DU to cluster-level signal processing. We notice that the fronthaul load in this paper only considers quantities directly related to combining and detection in the UL and the information bits in the DL. The coefficients needed to compute the weights $\{w_{\ell,k}\}$ and $\{w_{\ell,k}^0\}$ for UL combining and DL precoding, respectively, are not considered. Note that these coefficients are constant during a coherence block of T signal dimensions and that $T - \tau_p$ channel uses are dedicated to UL and DL data transmission. Assuming $\tau_p \approx 25$ and T typically in the hundreds or thousands (see [14]), the fronthaul load related to the $T - \tau_p$ channel uses for data transmission is relatively large compared to the number of coefficients needed to compute the UL combining and DL precoding weights (see [2]). We assume that these coefficients can be quantized very accurately without increasing significantly the fronthaul load. Therefore, we do not take into account the fronthaul traffic to exchange the coefficients for computing $\{w_{\ell,k}\}$ and $\{w_{\ell,k}^0\}$ when defining the UL and DL fronthaul load per RB.

III. PROBLEM STATEMENT

In this section, we will describe how the user PHY data rates translate to the UL and DL fronthaul load of the whole system. The network operates in time division duplex mode, where $\gamma_{\text{DL}} \in (0, 1)$ denotes the resource fraction allocated to the DL, and as a consequence $(1 - \gamma_{\text{DL}})$ is allocated to the UL. We use half-duplex fronthaul links as in [2], i.e., the fronthaul flow constraints must incorporate the fact that the data rate generated by the RUs in the UL is weighted by a factor $(1 - \gamma_{\text{DL}})$ and the data rate generated by the DUs in the DL is weighted by a factor γ_{DL} . The amount of UL fronthaul load transmitted from RU ℓ to router q , from router q to router q' , and from router q to DU n regarding user k is denoted as $x_k^{\text{ru}}(\ell, q)$, $x_k^{\text{fh}}(q, q')$, and $x_k^{\text{du}}(q, n)$, respectively. For the DL, we use a similar notation, where $y_k^{\text{ru}}(q, \ell)$, $y_k^{\text{fh}}(q, q')$ and $y_k^{\text{du}}(n, q)$ is the amount of DL fronthaul data with respect to user k sent from router q to RU ℓ , from router q to router q' , and from DU n to router q , respectively. The load of any fronthaul link $(a, b) \in \mathcal{E}_{\text{front}}$ is given by $\sum_k x_k(a, b)$, where $x_k(a, b)$ is one of the load variables previously introduced.

A. UL fronthaul load flow

As defined in (5), RU $\ell \in \mathcal{C}_k$ generates a quantization rate of $B_{\ell,k}$ bits per channel use that must be sent via the fronthaul to the DU n that is the cluster processor of the user-centric cluster \mathcal{C}_k . Since the UL fronthaul contains only unicast flows, for each user k every router q must satisfy

$$\sum_{\ell} x_k^{\text{ru}}(\ell, q) + \sum_{q''} x_k^{\text{fh}}(q'', q) = \sum_n x_k^{\text{du}}(q, n) + \sum_{q'} x_k^{\text{fh}}(q, q'). \quad (14)$$

Considering the TDD resource allocation fraction γ_{DL} , the UL fronthaul flow constraint for UE-RU pair (ℓ, k) is

$$\sum_q x_k^{\text{ru}}(\ell, q) \geq a_{\ell,k}(1 - \gamma_{\text{DL}})B_{\ell,k}, \quad \forall k, \ell, \quad (15)$$

$$\text{and } x_k^{\text{ru}}(\ell, q) \leq a_{\ell,k}(1 - \gamma_{\text{DL}})B_{\ell,k}, \quad \forall k, \ell, q, \quad (16)$$

where the UE-RU association binary variable $a_{\ell,k} = 1$ if $(\ell, k) \in \mathcal{E}_{\text{ran}}$, and $a_{\ell,k} = 0$ if $(\ell, k) \notin \mathcal{E}_{\text{ran}}$. Now, let $b_{k,n} \in \{0, 1\}$ denote the cluster processor placement variable, defined by

$$b_{k,n} = \begin{cases} 1, & \text{if } \mathcal{C}_k \text{ is hosted by DU } n, \\ 0, & \text{otherwise.} \end{cases}$$

We have the constraints that each \mathcal{C}_k must be hosted by exactly one DU and a computation capacity constraint per DU, i.e.,

$$\sum_{n=1}^N b_{k,n} = 1, \quad \forall k, \quad (17)$$

$$\sum_{k=1}^K b_{k,n} \leq Z_n, \quad \forall n, \quad (18)$$

where Z_n is a computation limit for the number of cluster processors at any DU n . Given the cluster processor placement, we note that the received flow relative to user k to DU n hosting \mathcal{C}_k must be not smaller than the sum of source rates $(1 - \gamma_{\text{DL}})B_{\ell,k}$ over all RUs $\ell \in \mathcal{C}_k$, i.e.,

$$\sum_q x_k^{\text{du}}(q, n) \geq b_{k,n}(1 - \gamma_{\text{DL}}) \sum_{\ell \in \mathcal{C}_k} B_{\ell,k}, \quad \forall k, n. \quad (19)$$

We also introduce the individual load variables upper bounds

$$x_k^{\text{du}}(q, n) \leq b_{k,n}(1 - \gamma_{\text{DL}}) \sum_{\ell \in \mathcal{C}_k} B_{\ell,k}, \quad \forall k, q, n. \quad (20)$$

In particular, this means that if $b_{k,n} = 0$, the rate relative to user k from any connected router q to n will be zero.

B. DL fronthaul load flow

As explained in Section II-C, the DL fronthaul traffic is of type multiple multicast and the number of information bits per channel use necessary to encode the DL signal for user k at each RU $\ell \in \mathcal{C}_k$ is equal to the DL PHY rate R_k^{dl} . Since the DL fronthaul traffic is of multicast type, the flow conservation at the routers no longer applies (e.g., intermediate nodes may duplicate some input to multiple output links). Considering the data of user k , the output data of a router q to any RU ℓ or

router q'' must be less than or equal to the sum input data from the cluster processor and other routers, i.e.,

$$\sum_n y_k^{\text{du}}(n, q) + \sum_{q'} y_k^{\text{fh}}(q', q) \geq y_k^{\text{ru}}(q, \ell), \quad \forall k, q, \ell, \quad (21)$$

and

$$\sum_n y_k^{\text{du}}(n, q) + \sum_{q'} y_k^{\text{fh}}(q', q) \geq y_k^{\text{fh}}(q, q''), \quad \forall k, q, q''. \quad (22)$$

In the DL, the DU hosting C_k must transmit at least R_k^{dl} bits per channel use to the connected routers, i.e.,

$$\sum_q y_k^{\text{du}}(n, q) \geq b_{k,n} \gamma_{\text{DL}} R_k^{\text{dl}}, \quad \forall k, n. \quad (23)$$

On each individual link to a router q , the DU needs to transmit at most R_k^{dl} bits and only the DU hosting cluster C_k transmits fronthaul data for user k . These two constraints are summarized as

$$y_k^{\text{du}}(n, q) \leq b_{k,n} \gamma_{\text{DL}} R_k^{\text{dl}}, \quad \forall k, n, q. \quad (24)$$

The constraint that guarantees that each RU $\ell \in C_k$ receives R_k^{dl} fronthaul bits for user k is formulated as

$$\sum_q y_k^{\text{ru}}(q, \ell) \geq a_{k,\ell} \gamma_{\text{DL}} R_k^{\text{dl}}, \quad \forall k, \ell. \quad (25)$$

C. Fronthaul optimization problem

Let $\mathcal{C} = \{C_L, C_Q, C_D\}$ the maximum link loads for RU-router, router-router, and router-DU links (with corresponding weights $\eta_L/\eta_Q/\eta_D$). Further, \mathcal{B} , \mathcal{X} and \mathcal{Y} denote the ensemble of all $\{b_{k,n} : \forall k, n\}$, UL load and DL load variables, respectively. Then, we can use the mixed-integer linear program (MILP) of [2] for joint optimization:

$$\min_{\mathcal{B}, \mathcal{C}, \mathcal{X}, \mathcal{Y}} \quad \eta_L C_L + \eta_Q C_Q + \eta_D C_D \quad (26a)$$

$$\text{s. t.} \quad \sum_k (x_k^{\text{ru}}(\ell, q) + y_k^{\text{ru}}(q, \ell)) \leq C_L, \quad \forall \ell, q, \quad (26b)$$

$$\sum_k (x_k^{\text{fh}}(q, q') + y_k^{\text{fh}}(q, q')) \leq C_Q, \quad \forall q, q', \quad (26c)$$

$$\sum_k (x_k^{\text{du}}(q, n) + y_k^{\text{du}}(n, q)) \leq C_D, \quad \forall q, n, \quad (26d)$$

$$(14) - (25), \quad (26e)$$

where constraints (26b)-(26d) ensure that each link load is less than the link capacity. Note that MILPs can be solved by existing highly efficient optimization tools [15]. To implement (26), we utilize the MATLAB function `intlinprog`.

IV. NUMERICAL RESULTS

We start this section with an overview of the considered network and parameters, before showing the impact of different K and D on the PHY SE and fronthaul load. Our cell-free massive MIMO network has an area of $200 \times 200 \text{ m}^2$ with a torus topology to avoid boundary effects. The LSFCs are given by the 3GPP urban microcell pathloss model from [16, Table 7.4.2-1]. The spatial correlation between antennas follows the

one-ring scattering model with angular support $\Theta_{\ell,k} = [\theta_{\ell,k} - \Delta/2, \theta_{\ell,k} + \Delta/2]$, i.e., it is centered at angle $\theta_{\ell,k}$ of the LOS between RU ℓ and UE k with angular spread Δ . Then, the channel between UE k and RU ℓ is given by

$$\mathbf{h}_{\ell,k} = \sqrt{\frac{\beta_{\ell,k} M}{|\mathcal{S}_{\ell,k}|}} \mathbf{F}_{\ell,k} \boldsymbol{\nu}_{\ell,k},$$

where the index set $\mathcal{S}_{\ell,k} \subseteq \{0, \dots, M-1\}$ includes all integers m such that $2\pi m/M \in \Theta_{\ell,k}$ (where angles are taken modulo 2π), $\mathbf{F}_{\ell,k}$ is the submatrix extracted from the $M \times M$ unitary DFT matrix \mathbf{F} by taking the columns indexed by $\mathcal{S}_{\ell,k}$, and $\boldsymbol{\nu}_{\ell,k} \in \mathbb{C}^{|\mathcal{S}_{\ell,k}| \times 1}$ has i.i.d. components $\sim \mathcal{CN}(0, 1)$. Hence, $\mathbf{h}_{\ell,k}$ is a Gaussian zero-mean random vector confined in the subspace spanned by the columns of $\mathbf{F}_{\ell,k}$.

A. Simulation setup

We consider a system with $L = 20$ RUs, each equipped with $M = 10$ antennas and placed on a rectangular 5×4 grid. We set $\Delta = \pi/8$, $\tau_p = 25$, the maximum cluster size $|C_k| \leq C_{\max} = 7$, and the SNR threshold $\eta = 1$ in (1). We define the expected pathloss $\bar{\beta}$ at distance $2.5d_L$ between UE and RU, where $d_L = \sqrt{\frac{A}{\pi L}}$ is the radius of a disk of area equal to A/L . The UL energy per symbol of each UE is $\bar{\beta} M \text{SNR} = 1$ (i.e., 0 dB) and thus depends on the RU density and the number of RU antennas. This choice leads to a certain level of coverage overlap among RUs, such that each UE is likely to associate with several RUs. The complex baseband noise power density is $N_0 = -174 \text{ dBm/Hz}$. The RUs are partially connected to $Q = 5$ routers, which in turn are partially connected to $N = 4$ DUs.³ For each of the DUs, we impose $Z_n = K/2$. The optimization weights η_L, η_Q, η_D are set to 1.

Each coherence block contains $T = 200$ signal dimensions, of which $\tau_p = 20$ are used for UL pilots and $\gamma_{\text{DL}} = 0.8$. We compute the expectation in (10) and (13) over 100 channel realizations and define the total UL/DL SE, i.e., the UL/DL PHY SE in bits per channel use (or bit/s/Hz) of all users, as $\text{SE}^{\text{ul}} = (1 - \gamma_{\text{DL}}) (1 - \frac{\tau_p}{T}) \sum_{k \in \mathcal{K}} R_k^{\text{ul}}$ and $\text{SE}^{\text{dl}} = \gamma_{\text{DL}} (1 - \frac{\tau_p}{T}) \sum_{k \in \mathcal{K}} R_k^{\text{dl}}$ bit/s/Hz. The total sum PHY SE is given by $\text{SE}_{\text{tot}} = \text{SE}^{\text{ul}} + \text{SE}^{\text{dl}}$.

B. Evaluation of the user load and quantization level

We evaluate the impact of the user load K and distortion level D on the fronthaul load and PHY SE. We let $K = [75, 200]$, which ranges from a lightly loaded to an overloaded system compared to LM (see [17]). Recall that a UE-RU association is removed if $\sigma_{\ell,k}^2 < D$. Therefore, the smallest distortion level D is chosen such that $D/\sigma_{\min}^2 < 1$, where $\sigma_{\min}^2 = \min_{(\ell,k) \in \mathcal{E}_{\text{ran}}} \sigma_{\ell,k}^2$. As D is increased, $B_{\ell,k}$ in (5) decreases and more UE-RU associations are removed.

As Fig. 2 shows, each increase of D leads to a significant reduction of the fronthaul load, i.e., $\eta_L C_L + \eta_Q C_Q + \eta_D C_D$, while the sum PHY SE only decreases for large D . We note

³We use the same fronthaul links as [2] and refer the reader to [2] for a description and illustration of what links exist between RUs, routers and DUs.

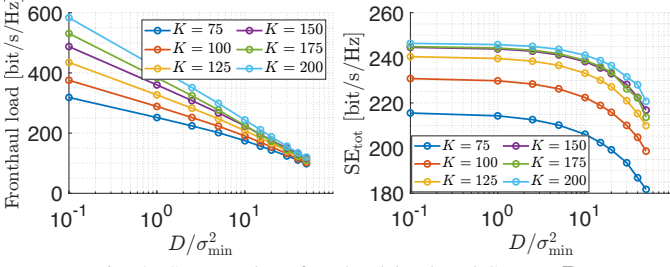


Fig. 2: Sum UL/DL fronthaul load and SE vs. D .

that as D is increased, the fronthaul load becomes smaller for large K than for small K in some cases. This is explained by smaller $\{\sigma_{\ell,k}^2 : (\ell, k)\}$ for large K due to multi-user interference. Increasing D leads to more UE-RU associations being removed and reduced quantization rates $B_{\ell,k}$ compared to a low user load. When D is increased in the high distortion regime, the sum PHY SE first decreases slightly and then, as more UE-RU associations are removed, significantly. Fig. 3 confirms and explains this behavior. The degradation of the 5th percentile DL user SE $\gamma_{\text{DL}} (1 - \frac{\tau_p}{T}) R_k^{\text{dl}}$ is observed at $D/\sigma_{\min}^2 \geq 10$. No quantization is needed in the DL, so this effect is caused solely by removing UE-RU associations. The degradation of $(1 - \gamma_{\text{DL}}) (1 - \frac{\tau_p}{T}) R_k^{\text{ul}}$ in the UL occurs for smaller D due to fronthaul quantization distortion. Since a larger fraction of resources is used in the DL, the impact of D on SE_{tot} in Fig. 2 becomes more significant for $D/\sigma_{\min}^2 \geq 10$. We further notice that the average cluster size decreases from nearly $C_{\max} = 7$ RUs with $D/\sigma_{\min}^2 = 1$ to ≈ 6 and ≈ 5 RUs at $D/\sigma_{\min}^2 = 10$ and $D/\sigma_{\min}^2 = 20$, respectively. As another result from Fig. 2 we observe that the sum SE grows with larger K , but only in the range $K = [75, 150]$. For $K > 150$, multi-user interference becomes more severe and it is recommended not to serve much more than 150 UEs.

C. Concluding remarks and future work

A wide range of user loads can be supported by the same cell-free massive MIMO RAN and fronthaul network if the distortion level D is optimized. For example, in this setup, $D/\sigma_{\min}^2 \approx 5$ for $K = \{75, 100\}$ and $D/\sigma_{\min}^2 \approx 10$ for $K = \{125, 150\}$ are good choices. Then, the fronthaul load for the recommended user load $K = [75, 150]$ is between 200 and 225 bit/s/Hz (i.e., in a relatively small margin) and the PHY SE is degraded only very slightly compared to smaller D . This is also an interesting result for scheduling problems. If the total number of users K_{tot} in the network area is varying, the scheduler can flexibly react to different user loads and demands, and change the number K of simultaneously active users without overloading the fronthaul network. We conclude that dense cell-free massive MIMO deployments are a very promising approach to achieve resilient 6G (radio access and fronthaul) networks with regard to varying user densities.

Future work includes investigating the impact of the distortion level for different fronthaul network topologies and fronthaul load optimization algorithms, and to study scheduling with flexible user loads under fronthaul constraints.

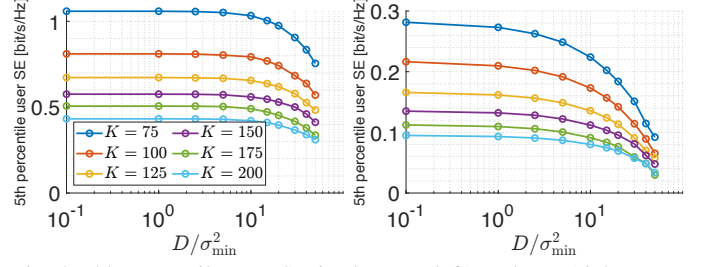


Fig. 3: 5th percentile user SE in the DL (left) and UL (right, same legend as for the UL applies) vs. D .

V. ACKNOWLEDGMENTS

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany in the programme of “Souverän. Digital. Vernetzt.” Joint project 6G-RIC, project identification number 16KISK030.

REFERENCES

- [1] H. Q. Ngo *et al.*, “Ultradense Cell-Free Massive MIMO for 6G: Technical Overview and Open Questions,” *Proceedings of the IEEE*, 2024.
- [2] Z. Li *et al.*, “Joint Fronthaul Load Balancing and Computation Resource Allocation in Cell-Free User-Centric Massive MIMO Networks,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 14 125–14 139, 2024.
- [3] A. Joshi *et al.*, “Fronthaul Resource Optimization for Cell-Free Massive MIMO Networks,” in *2024 14th International Workshop on Resilient Networks Design and Modeling (RNDM)*, to appear. IEEE, 2024.
- [4] E. Björnson *et al.*, “Scalable Cell-Free Massive MIMO Systems,” *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.
- [5] I. Chih-Lin *et al.*, “Rethink fronthaul for soft RAN,” *IEEE Communications Magazine*, vol. 53, no. 9, pp. 82–88, 2015.
- [6] R. B. Dilmaghani *et al.*, “On designing communication networks for emergency situations,” in *2006 IEEE International Symposium on Technology and Society*. IEEE, 2006, pp. 1–8.
- [7] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.
- [8] N. Osawa *et al.*, “Overloaded Pilot Assignment with Pilot Decontamination for Cell-Free Systems,” in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.
- [9] F. Götsch *et al.*, “Subspace-Based Pilot Decontamination in User-Centric Scalable Cell-Free Wireless Networks,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 6, pp. 4117–4131, 2023.
- [10] A. Adhikary *et al.*, “Joint Spatial Division and Multiplexing—The Large-Scale Array Regime,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.
- [11] Ericsson, “Driving Open RAN forward – An improved open fronthaul interface bringing performance to Open RAN,” Tech. Rep., 2023.
- [12] G. Caire, “On the Ergodic Rate Lower Bounds With Applications to Massive MIMO,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3258–3268, 2018.
- [13] ITU-T G-series Recommendations – Supplement 66, “5G Wireless Fronthaul Requirements in a Passive Optical Network Context,” Geneva, Switzerland, Sep 2020.
- [14] R. P. Torres *et al.*, “A lower bound for the coherence block length in mobile radio channels,” *Electronics*, vol. 10, no. 4, p. 398, 2021.
- [15] E. El Haber *et al.*, “Joint Optimization of Computational Cost and Devices Energy for Task Offloading in Multi-Tier Edge-Clouds,” *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3407–3421, 2019.
- [16] 3GPP, “Study on channel model for frequencies from 0.5 to 100 GHz (Release 16),” 3GPP Tech. Spec. 38.901, 12 2019, Version 16.1.0.
- [17] F. Götsch *et al.*, “Fairness Scheduling in User-Centric Cell-Free Massive MIMO Wireless Networks,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 11 942–11 957, 2024.